The Basics of RDM

The principles, the hows, and the whys

Felix Rau
(University of Cologne)

Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

The plan for today

- RDM & data
- Principles of good RDM
- Concrete aspects of RDM

Research Data Management

Research Data Management is a comprehensive set of practices that enable researchers to acquire, store, handle, organize, preserve, share, and publish their data effectively and responsibly throughout the research process.

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

Research Data Management

is

acquiring

keeping

handling

releasing

data in controlled and documented workflows.

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

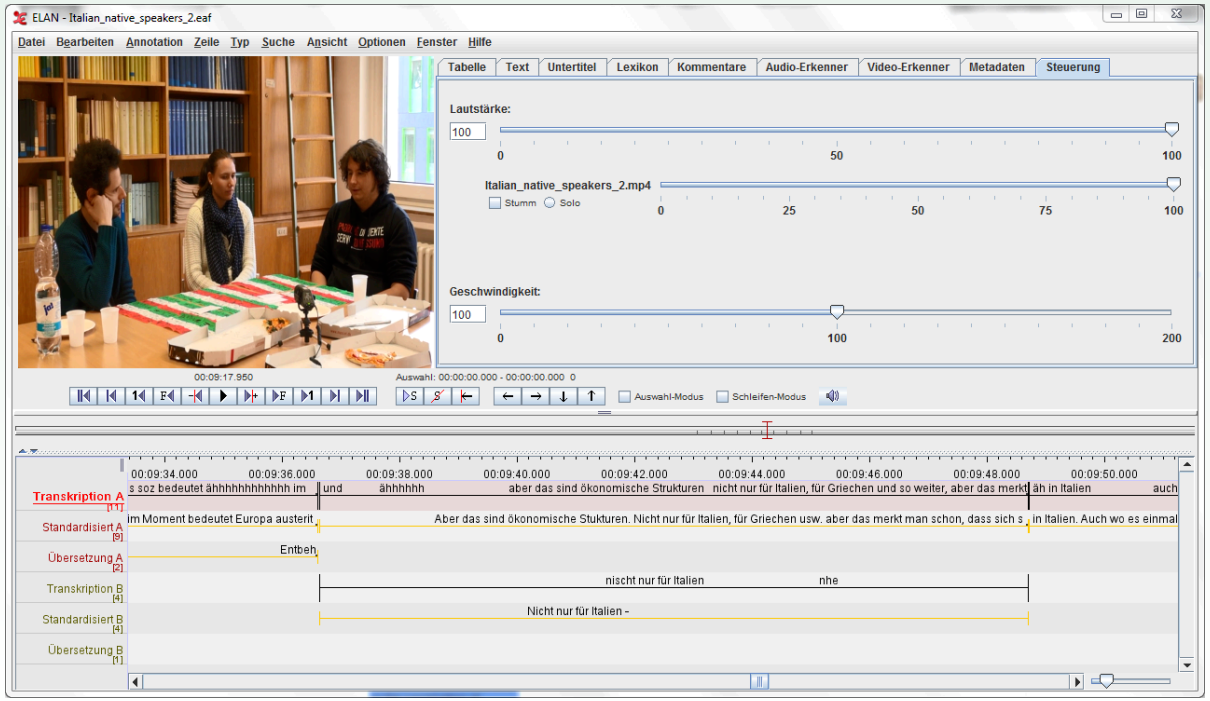
Data

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

How does your data look like?

Data

In linguistics: Representation of a spoken, signed, or written speech event.



Data types

texts 📚, audio 🎤, datenbases 📀, videos 📽, eye tracking 👀, tables 📊, manuscripts 📝, software 💾, fmri data 🦴, ...

Data

Leonelli 2016 defines data as

> any product of research activities, ranging from artifacts such as photographs to symbols such as letters or numbers, that is collected, stored, and disseminated in order to be used as evidence for knowledge claims.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study.* Chicago ; London: The University of Chicago Press.

Data models

interatcional speech event → video recording → transcript → translation → morpheme gloss → annotation of intonation

30 recording 👆 ⇒ numerical Datenset in a table

How about you? 

What else is there?

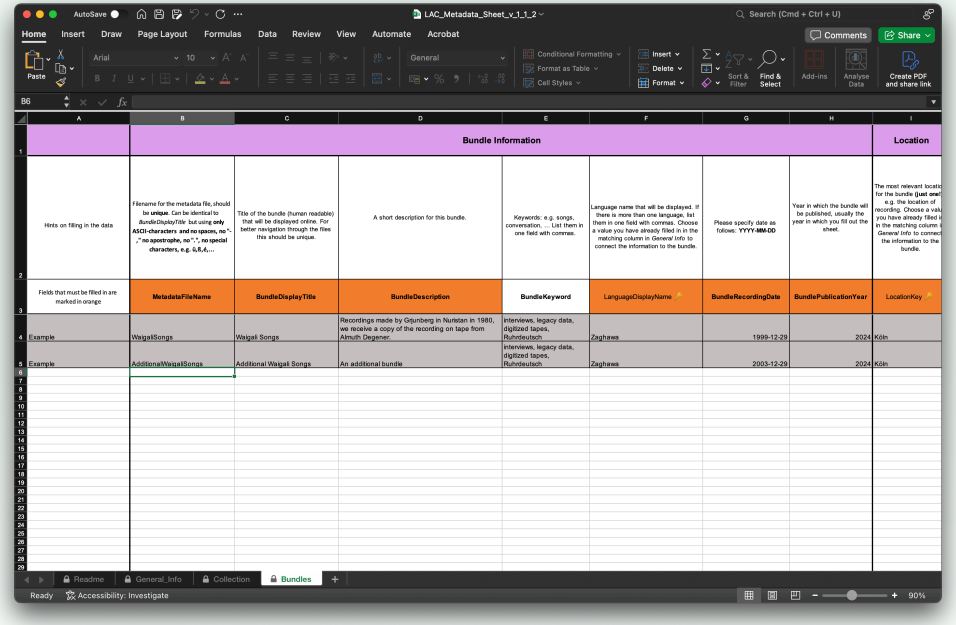
Metadata

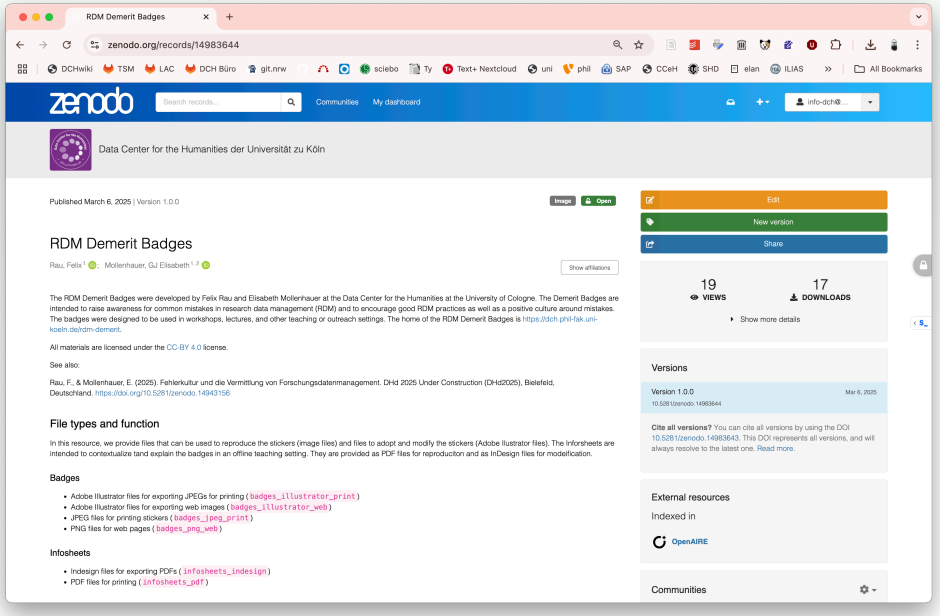
DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

Metadata

Metadata is structured information about research data that enables and supports the collection, storage, analysis and dissemination of data.

Metadata


LAC_Metadata_Sheet_v_1_1_2
Bundle Information
Location
MetadataFileName
BundleDisplayTitle
BundleDescription
BundleKeyword
LanguageDisplayName
BundleRecordingDate
BundlePublicationYear
LocationKey
WaigaliSongs
Waigali Songs
Recordings made by Gjunberg in Nuristan in 1980, we receive a copy of the recording on tape from Almuth Degener.
Interviews, legacy data, digitized tapes, Ruhrdeutsch
Zaghawa
1999-12-29
2024
Köln
AdditionalWaigaliSongs
Additional Waigali Songs
An additional bundle
Interviews, legacy data, digitized tapes, Ruhrdeutsch
Zaghawa
2003-12-29
2024
Köln
Readme
General_Info
Collection
Bundles


RDM Demerit Badges
zenodo.org/records/14983644
Data Center for the Humanities der Universität zu Köln
Published March 6, 2025 | Version 1.0.0
RDM Demerit Badges
Rau, Felix; Mollenhauer, GJ Elisabeth
The RDM Demerit Badges were developed by Felix Rau and Elisabeth Mollenhauer at the Data Center for the Humanities at the University of Cologne. The Demerit Badges are intended to raise awareness for common mistakes in research data management (RDM) and to encourage good RDM practices as well as a positive culture around mistakes. The badges were designed to be used in workshops, lectures, and other teaching or outreach settings. The home of the RDM Demerit Badges is https://dch.phil-fak.uni-koeln.de/rdm-demerit.
All materials are licensed under the CC-BY 4.0 license.
See also:
Rau, F., & Mollenhauer, E. (2025). Fehlerkultur und die Vermittlung von Forschungsdatenmanagement. DHd 2025 Under Construction (DHd2025), Bielefeld, Deutschland. https://doi.org/10.5281/zenodo.14943156
File types and function
In this resource, we provide files that can be used to reproduce the stickers (image files) and files to adopt and modify the stickers (Adobe Illustrator files). The Infosheets are intended to contextualize tand explain the badges in an offline teaching setting. They are provided as PDF files for reproduciton and as InDesign files for modification.
Badges
Adobe Illustrator files for exporting JPEGs for printing (badges_illustrator_print)
Adobe Illustrator files for exporting web images (badges_illustrator_web)
JPEG files for printing stickers (badges_jpeg_print)
PNG files for web pages (badges_png_web)
Infosheets
Indesign files for exporting PDFs (infosheets_indesign)
PDF files for printing (infosheets_pdf)
Edit
New version
Share
19 VIEWS
17 DOWNLOADS
Show more details
Versions
Version 1.0.0
Mar 6, 2025
Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.14983643. This DOI represents all versions, and will always resolve to the latest one. Read more.
External resources
Indexed in
OpenAIRE
Communities

Paradata

> The paradata of a data set or survey are data about the process by which the data were collected. Wikipedia

What do we have to manage?

- Data
- Various data models
- Metadata
- Paradata

Let's start with the basics of RDM

Principles of good RDM

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

FAIR

https://www.go-fair.org/fair-principles/

- Findable
- Accessible
- Interoperable
- Reusable

- F1: (Meta) data are assigned globally unique and persistent identifiers
- A1: (Meta)data are retrievable by their identifier using a standardised communication protocol

Practical principles

- Plan your RDM and revise it
- Document structures
- Document decisions
- Document changes (e.g. through code)
- Version your data
- Automatize processes
- Make processing repeatable
- Prioritize control

Stages of RDM

Data management plans

- Planning is extremely important
- DMPs are a way to start planning
- They are not a magic bullet
- They can be an administrative requirement
- *Have a plan and write it down!*

- acquiring
- keeping
- handling
- releasing

Acquiring

- Document data and how data was acquired
  - Metadata
  - Paradata
- Data quality
- Legal aspects
- Ethical aspects
- Document data sensitivity
- Document consent

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

Keeping

- Storage
- Backup
- Data/File typen
- Folder structures
- File names
- Data security

Storage

- Field vs Office
- Water/dust resistance (IP65)
- University of Cologne
  - SOFS
  - RDS
  - Sciebo

Rau, F. (2023). *DCH Project Data Storage Guidelines* (1.0.0). Zenodo. doi:10.5281/zenodo.7760967

Rau, F. (2023). *DCH Field Data Storage Guidelines* (1.0.0). Zenodo. doi:10.5281/zenodo.7957791

Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

DCH Field Data Storage Guidelines

Version 1.0.0

1. Keep three instances
2. Keep two local instances on different media
3. Keep one off-site copy
4. Automatize data replication and backup
5. Document procedure and responsibilities

1 Three instances

Keep three instances of the data. Besides the original working instance, keep two copies of your project data.

2 Two local instances on different media

Use dust- and waterproof storage media (at least protected according to IP55, better IP67) for the local copy and the off-site copy, if the latter is not internet-based.

3 One off-site copy

If your field situation does not allow for an internet-based off-site copy, keep one copy as removed from the other copies as possible. This can mean keeping it in a separate bag or a separate room.

4 Automatize data replication and backup

The process of regular local replication and off-site backup should be as regular as possible, and automatized if possible. The frequency of replication and backup must be adjusted to the maximum acceptable amount of data loss in case of an incident.

5 Document procedure and responsibilities

Document the data storage procedure – including location of the three copies, frequency of replication and backup – as well as the responsibilities for implementing the procedure e.g. in your data management plan.

DCH Field Data Storage Guidelines. Version 1.0.0 10.5281/zenodo.7957791

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

Backup

- 3-2-1 Rule
  - 3 copies of your data
  - 2 different media
  - 1 off-site copy
- 3-2-1-1-0 Rule
  - 1 offline copy
  - 0 errors

Rau, F. (2023). *DCH Project Data Storage Guidelines* (1.0.0). Zenodo. doi:10.5281/zenodo.7760967

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

File formats

Openness	Standardised, patents, readability (Software)
Adoption	Market share or number of implementations
Complexity	Human readable
Technical Protection Mechanism	DRM
Self-documentation	Metadata, self-documenting
Robustness	Corruption detection, Backward/forward compatibility
Dependencies	Hardware, OS, Software dependency

Folder structure

```
project/
├── 01_admin/
├── 02_data/
│   ├── 01_raw/
│   └── 02_processed/
├── 03_code/
├── 04_docs/
├── 05_output/
├── README.md
└── LICENSE
```

Colomb, Julien, Thorsten Arendt, Keisuke Sehara, and The Gin-Tonic team. 2021. “Towards a Standardized Research Folder Structure.” Generation Research. Webarchive

GIN-Tonic: Research folder structure standard https://gin-tonic.netlify.app/standard/

Rau, F. (2023). *DCH Folder Structure Guidelines* (1.0.0). Zenodo.

doi:10.5281/zenodo.7452113

Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

DCH Folder Structure Guidelines

Version 1.0.0

1. Separate types of information
2. Separate stages of processing (e.g. raw, cleaned, annotated)
3. Keep the folder depth at 4–5 levels
4. Order folders with leading numbers in folder names
5. Document your folder structure

1 Separate types of information

Separate types of information, including project and administrative information.

Example:

```
project/
├─ 01_project_management/
├─ 02_data/
├─ 03_analysis/
└─ 04_publications/
```

2 Separate stages of processing

Separate different states of the data.

Example:

```
├─ 02_data/
│  ├─ 01_raw_data/
│  ├─ 02_cleaned_data/
│  └─ 03_annotated_data/
```

3 Keep the folder depth at 4–5 levels

Avoid too deeply nested folder hierarchies by keeping the depth to 4–5 levels.

4 Order folders with leading numbers in folder names

Begin folder names on all levels with padded numbers to facilitate sorting.

Example :

```
├─ 03_analysis/
│  ├─ 01_skript/
│  └─ 02_output/
```

5 Document your folder structure

Document your folder structure in a README file placed in the top folder of the dataset.

See: DCH Readme File Guidelines doi:10.5281/zenodo.7447616

DCH Folder Structure Guidelines. Version 1.0.0 doi:10.5281/zenodo.7452113

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

File naming

1. Give unambiguous, meaningful, readable, but succinct names
2. Chose names that are safe across file and operating systems
3. Structure the filename and use filename extensions
4. Facilitate alphabetical sorting
5. Document your naming pattern

Rau, F. (2023). DCH File Naming Guidelines

(1.0.0). Zenodo.

doi:10.5281/zenodo.7447485

The Basics of Resarch Data Management

DCH File Naming Guidelines

Version 1.0.0

1. Give unambiguous, meaningful, readable, but succinct names
2. Chose names that are safe across file and operating systems
3. Structure the filename and use filename extensions
4. Facilitate alphabetical sorting
5. Document your naming pattern

1 Give unambiguous, meaningful, readable, but succinct names

Select relevant characteristics of the file content as part of the names and use unambiguous labels for them. For example, use `en` and `it` (ISO 639-1) instead of English and Italian.

See: DCH File Naming Recommendations doi:0.5281/zenodo.7447562

2 Chose names that are safe across file and operating systems

Restrict the character inventory to the lowercase Latin alphabet `a-z`, digits `0-9`, hyphen (minus) -, and underscore _. Additionally, the full stop . is used once to separate the filename extension.

Example: `original-text_2009-04-23_001.mp4`

3 Structure the filename and use filename extensions

Separate parts of the name by underscore _ and structure parts with hyphen (minus) -.

Example: `en_session-01_section-a.mp4`

4 Facilitate alphabetical sorting

Pad numbers with zeros to facilitate accurate sorting. Format dates following the YYYY-MM-DD (ISO 8601) pattern.

Example : `speaker-01_2009-04-23_take-001.wav`

5 Document your naming pattern

Document your naming pattern, ideally in the same location where the files can be found. For example, place a README file in the top folder of the project or dataset.

See: DCH Readme File Guidelines doi:10.5281/zenodo.7447616

DCH File Naming Guidelines. Version 1.0.0 doi:10.5281/zenodo.7447485

DCH Data Center for the Humanities Kölner Datenzentrum für die Geisteswissenschaften

Examples file naming

`[ISO 639-3]_[speaker]_[date]_[number]_[data type].[file extension]`

```
ger_LK_2023-05-31_001_raw-audio.wav
ger_LK_2023-05-31_001_raw-video.m4v
```

Rau, F. (2023). DCH File Naming Guidelines (1.0.0). Zenodo. doi:10.5281/zenodo.7447485

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

Data security

- Access control
- Encryption in motion
- Encryption in storage
  - File encryption (7zip)
  - Folder encryption (VeraCrypt)
  - Disk encryption (most OS provide this)
  - Cloud storage encryption (Cryptomator)

Handling

- Versioning
- Everything with a script
- Automatize!
- (Analysis)

Versioning

- File based versioning (`filename_v2.pdf`)
- git
- DataLad
- DVC

Everything with a script

> The most basic aspect of reproducible research is that everything you do (convert data files, clean data, analyze data) should be accomplished via code. Pointing and clicking, and copy-paste, are not reproducible.
>
> Karl Broman Everything with a script

- Don’t hand-edit data files
- Data cleaning should be in scripts
- Analysis should be in scripts

Automatize!

> Ideally, the reproduction of your results is a one-button operation. And this is valuable not just for others, but also for yourself (or your future self). For example, if the primary data should change (and it often does), wouldn't it be nice to have one command that re-runs everything?
>
> Karl Broman Automate the process

- Run all your scripts from one script
- Use GnuMake or similar

Releasing

- Deleting
- Archiving
- Publishing

Deleting

- Not every thinhs can be preserved
- Not everything needs to be preserved
- **...but** often it should be preserved

Archiving

- Long-term preservation
- Continued viability
- Focus on preservation not on access

Publishing

- Data repositories
- Domain-specific repositories
- General-purpose repositories
- As open as possible
- As closed as necessary
- Data licenses

- Here the FAIR principles come into play again.
- Pick the most FAIR repository for your data.
- R1.3: (Meta)data meet domain-relevant community standards

Take away message

- Document, version, automatize
- Make processing repeatable
- Prioritize control

Data Center for the Humanities

The research data center of the Faculty of Arts and Humanities at the University of Cologne. If you have any questions about research data management, please feel free to contact us.

Danke!

info-dch@uni-koeln.de

Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften